**Forecasting disease burden in MDD**

**SCAN-2030 Work Package 2 [WP2] Analysis Plan**

## 1. Objective

The primary objective of this study is to forecast the incidence, prevalence, and healthcare burden of major depressive disorder (MDD) in Hong Kong over the next ten years (2023–2032). By leveraging historical data from 2014 to 2022, we aim to predict future trends using statistical modeling techniques. These forecasts will support healthcare planning, policy-making, and resource allocation to better address the growing needs of MDD patients.

## 2. Rationale

Understanding the future burden of MDD is crucial for effective healthcare decision-making. Forecasting trends in disease incidence and prevalence will help policymakers and healthcare providers anticipate future demands on the healthcare system. Additionally, estimating the economic impact and unmet treatment needs will aid in the development of innovative treatment strategies and healthcare policies. By applying real-world data and forecasting models, we can generate evidence-based insights to optimize mental health services and improve patient outcomes.

## 3. Data Source

The study will utilize data from the Clinical Data Analysis and Reporting System (CDARS), a comprehensive electronic medical record database managed by the Hospital Authority (HA) of Hong Kong. CDARS includes patient demographics, hospital attendance records, diagnoses, prescriptions, and laboratory test results from inpatient, outpatient, and emergency care settings. Additionally, mid-year population data from the Census and Statistics Department of Hong Kong will be used to calculate age-standardized rates and extrapolate future trends.

## 4. Study Population

Patients diagnosed with major depressive disorder (MDD) between 1 January 2014 and 31 December 2022 will be identified using the ICD-9-CM diagnostic codes (296.2, 296.3, 300.4,

and 311). These patients will be categorized based on their year of diagnosis to establish historical trends. Forecasting models will then use these trends to predict future disease burden from 2023 to 2032. Prevalence, incidence, and cost data were aggregated based on analyses performed as part of SCAN-2030 Work Package 1.

## 5. Statistical Analysis

To forecast future trends in major depressive disorder (MDD) incidence, prevalence, and healthcare burden, we employ an AutoRegressive Integrated Moving Average (ARIMA) modeling approach. ARIMA and regression with ARIMA errors are well-established frameworks for analyzing and predicting time-series data, especially when historical patterns, autocorrelations, and potential interventions must be accounted for. For each outcome, candidate models are automatically generated and evaluated using the auto.arima function from the R forecast package, which systematically explores a wide range of plausible ARIMA configurations. To address issues of non-normality and heteroskedasticity commonly observed in healthcare and cost data, both the raw and log-transformed versions of each outcome series are modeled in parallel.

Our modeling strategy is enhanced through the incorporation of exogenous regressors. Binary indicator variables—such as those representing the onset of the COVID-19 pandemic (dum_pan) or significant social events (dum_so)—are included as external covariates (xreg) within the ARIMA framework. These covariates are intended to capture abrupt, non-repeating shifts in the time series that are not explained by endogenous patterns alone. For the historical period (2014–2022), they reflect actual events; for future forecasts (2023–2032), the default assumption is that these events do not recur, and the covariate values are set to zero unless otherwise justified.

### 5.1 Model Comparison and Selection Criteria

Model selection is guided by a combination of information criteria and error minimization. For each outcome and transformation, ARIMA models are fit to minimize both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), providing a balance

between model fit and complexity. All candidate models are extracted and ranked according to their information criterion scores. When the top two models are closely matched (i.e., score gap <2), the model with the higher sum of ARIMA orders (p+d+q) is preferred, as it offers greater flexibility in capturing temporal dependencies and trends—provided that this does not introduce overfitting. If model comparison is inconclusive or model refitting fails, the best default model from auto.arima is retained.

Comprehensive model validation is performed for each selected candidate. Residuals are inspected for randomness and absence of autocorrelation, and the mean squared error (MSE) is computed on the training data. For log-transformed models, fitted values are back-transformed to the original scale prior to MSE calculation to ensure comparability across modeling strategies. The final model selection for each outcome is based on the lowest observed MSE, either on the original or back-transformed scale, thereby prioritizing predictive accuracy. Additionally, historical back-testing is conducted by comparing model predictions within the observed period to actual values, providing a practical check on the model's forecasting fidelity.

**5.2 Forecasting Procedure**

After selecting the optimal ARIMA model for each outcome series, forecasts are generated for each year from 2023 to 2032. For models trained on log-transformed data, forecasts are exponentiated to yield interpretable predictions in the natural scale. Each forecast is accompanied by a 95% prediction interval, reflecting statistical uncertainty and enabling risk-informed planning. Forecasting is performed not only for aggregate outcomes but also separately for sex-specific and age-standardized incidence and prevalence, as well as for discrete components of healthcare costs (e.g., Accident and Emergency, inpatient, and outpatient services).

The results are systematically organized and exported in tabular form, with historical values (2014–2022) and forecasted values (2023–2032) presented side by side. This unified format facilitates interpretation, reporting, and downstream use in policy analysis or resource planning.

## 6. Study Outcomes

The study will focus on forecasting several key outcomes. Incidence forecasting will estimate the annual number of new MDD cases, while prevalence forecasting will predict the number of active patients seeking treatment. Age-standardized rates will be calculated to adjust for demographic changes. Additionally, the study will forecast healthcare resource utilization, including projected hospital admissions, outpatient visits, and inpatient length of stay. The economic burden of MDD will also be estimated by forecasting total healthcare costs associated with psychiatric care.